一、哈希碰撞概率

哈希碰撞的概率取决于两个因素(假设哈希函数是可靠的,每个值的生成概率都相同)

```
取值空间的大小(即哈希值的长度)
整个生命周期中,哈希值的计算次数
```

这个问题在数学上早有原型,叫做"生日问题"(birthday problem): 一个班级至少两个同学生日相同的概率?

答案很出人意料,如果至少两个同学生日相同的概率不超过5%,那么这个班只能有7个人;一个23人的班级有50%的概率,至少两个同学生日相同;50人班级有97%的概率

这意味着,如果哈希值的取值空间是365,只要计算23个哈希值,就有50%的可能产生碰撞

二、数学推导

至少两个人生日相同的概率,可以先算出所有人生日互不相同的概率,再用1减去这个概率

我们把这个问题设想成,每个人排队依次进入一个房间,第一个进入房间的人,与房间里已有的人(0人),生日都不相同的概率是 365/365 ,第二个进入房间的人，生日独一无二的概率是 364/365 ,第三个人是 363/365 ,以此类推

因此,所有人的生日都不相同的概率,就是下面的公式

$$\bar{p}(n) = 1 \cdot \left(1 - \frac{1}{365}\right) \cdot \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right)$$

上面公式的n表示进入房间的人数,可以看出进入房间的人越多,生日互不相同的概率就越小,那么至少有两个人生日相同的概率,就是1减去上面的公式

$$p(n) = 1 - \bar{p}(n) = 1 - \frac{365!}{365^n (365-n)!}$$

如果x是一个极小的值, e^x≈1+x ,因此生日问题的概率公式变成下面这样

$$\begin{aligned}\bar{p}(n) &\approx 1 \cdot e^{-\frac{1}{365}} \cdot e^{-\frac{2}{365}} \cdot e^{-\frac{n-1}{365}} \\ &= e^{-\frac{1+2+\cdots+(n-1)}{365}} \\ &= e^{-\frac{n(n-1)/2}{365}} = e^{-\frac{n(n-1)}{730}}.\end{aligned}$$

$$p(n) = 1 - \bar{p}(n) \approx 1 - e^{-\frac{n(n-1)}{730}}.$$

假设 d 为取值空间(生日问题里是365),就得到了一般化公式(哈希碰撞概率公式)

$$p(n,d) \approx 1 - e^{\frac{-n(n-1)}{2d}}$$

```
import  math
p=lambda  d,n:1-math.e**((1-n)*n/2/d)
p(365,23) #  0.5000017521827107
p(365,50) #  0.9651312540863107
```

三、应用

现在有一家公司,它的API每秒会收到100万个请求,每个请求都会生成一个哈希值,假定这个API会使用10年,能够接受的哈希碰撞概率是1000亿分之一(即每天发生一次哈希碰撞),假设哈希字符串由[0-9a-f]构成,请问哈希字符串最少需要多少个字符? (答案33个)

```
import  math

n=3600*24*365*10*1000000
p=1e-11
flag=10+6

d=lambda  p,n:n*(1-n)/math.log(1-p)/2   #  flag**length
length=math.log(d(p,n),flag)            #  32.967298736054715
```